

DOCUMENT RESUME

ED 266 688

FL 015 544

AUTHOR Lindblad, Torsten
TITLE On the Whys and Hows of Language Testing.
PUB DATE 85
NOTE 9p.; In: Language Testing in School. AFinLA Yearbook 1985. No. 41 (see FL 015 537). Document is in small print.
PUB TYPE Viewpoints (120)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Communicative Competence (Languages); Educational Objectives; Elementary Secondary Education; Foreign Countries; *Language Skills; *Language Tests; *Second Language Instruction; Standardized Tests; *Test Construction; *Test Use; *Test Validity
IDENTIFIERS Sweden

ABSTRACT

Six basic questions that are important to consider in foreign language testing are often taken for granted: (1) What is it that a student "knows" when he or she has acquired a foreign language, and is it measured in qualitative or quantitative terms or both? (2) What are the roles of skills and attitudes in learning a foreign language? (3) What implicit or explicit language norm is being taught? (4) What is the relationship between linguistic and communicative competence? (5) Is the view of language testing held by the teacher expansive or restrictive? and (6) Is the student the object of teaching efforts or the subject responsible for learning the language? Reflection on the questions indicates that tests can be given for a variety of purposes: diagnosis, determination of resource use, selection, guidance, examination, grading, evaluation, influence, stimulation, and research. Tests having different objectives are not generally interchangeable, a fact that is not often understood. Three Swedish test batteries having different objectives and applications (Swedish as a second language, standardized tests for grading, and free written production) illustrate the importance of analyzing and carefully relating the testing "whys" and "hows." (MSE)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

J. Torstén

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ON THE WHYS AND HOWS OF LANGUAGE TESTING

Torsten Lindblad (University of Gothenburg)

These reflections on why we test and how we test our students' command of a foreign language are not meant to be the all-embracing treatment on the somewhat grand scale that the title of the paper might suggest. Hopefully it will be, however, a reminder of the fact that our tests must never become ends in themselves but must always remain means to different ends. This is, at bottom, one way of approaching the problem of the validity of foreign-language tests.

1. Background

When people disagree in their opinions as to what is or what is not a good language test, the "deep structure" of the disagreement might be found in differences concerning certain underlying, basic concepts rather than in the "surface structure" of test format and item content. Let us, therefore, briefly mention six background concepts which might be useful in starting a discussion of what is often, implicitly, taken for granted. For reasons of space I shall have to limit myself to hinting at what is meant rather than treating in full each point mentioned.

1. View of knowledge. What is it a student "knows" when he has acquired command of a foreign language? What can be measured in quantitative terms and what must be estimated in qualitative ways? Is there a risk that too much attention is paid to what can be calculated in figures, as means and standard deviations etc? Also: If we say - as is often done - that a student knows a certain number of words in the foreign language, what exactly do we mean by "know" and what is "one word"? Is 'have' one word or three words in 'I have a dog', 'He will have it that' and 'You've been had'? And furthermore: In formulating the goals for the weakest students, i.e. in deciding on the basics of a foreign language, do we use a list of one hundred frequent words or do we mention a dozen important situations in which the student is supposed to be able to cope?

2. View of language teaching and learning. What does it mean to learn a foreign language? What is the role of the four skills as compared to attitudes to the foreign country and to speakers of the language and, for that matter, to speakers of foreign languages in general? And what attitudes to themselves and their own ability do our students acquire? Will they dare and be willing to use the foreign language when the situation arises? Is it, in other words, only new words and grammar paradigms that count?

3. View of language. What language do we teach, what language norm do we have, implicitly or explicitly? Where do we "draw the line" in accepting student responses in open-ended tests? Which of the following three sentences could be given points and what principles could be

formulated in order not to leave us in an ad hoc swamp: 'There's lots more questions I'd like to ask', 'There ain't gonna be any room for us', 'He said he don't like it'. I would suggest that comprehensibility is not enough (third example not accepted), usage will have to depend on the context (second example accepted in dialogue but not in written discourse) and suitability, a rather strict demand, well in accord with the concept of 'communicative competence', will have to be judged with the help of native informants (the first example fully acceptable today, whatever our grammar books might still say). For our purposes we might do well to use a prescriptive view in language teaching but a rather more descriptive one in testing. (These ideas are worked out in more detail in a paper, "The Language Norm - Does it Exist in English", by Professor Jan Svartvik, which can be ordered from the author of this paper.)

4. View of language skills. What is the relationship between linguistic competence and communicative competence? What are the relationships between the various factors of the foreign-language proficiency of the students? This is, of course, closely related to the view of knowledge and the problems mentioned under that heading.

5. View of language testing. Do we have (with Stevenson's terms) a restricted or an expanded view of language tests and testing? What are the risks in using one kind of test or another? Do the three well-known trends formulated by Spolsky - originally referred to as prescientific, structuralist-psychometric, and communicative - reflect swings of the pendulum between what C.P. Snow called "the two cultures"? If so, do we favour a communication-oriented, humanistic approach that "dares" to take into account aspects of language proficiency that cannot be objectively scored? How do we strike the happy medium between quantitative and qualitative approaches? Moreover: What do we want to test - what the students know, what they do not know, or what they should know?

6. View of students (human beings in general). Is the student the object of our teaching efforts or the subject responsible for learning the language? Do all students have the same ability to learn a foreign language? If not, does that influence our view of their "value" in any respect? Do all students have the same right to have the feeling that they have succeeded in some way? How do we handle the affective effects of our tests?

The way we answer all of the above questions will influence our views on the "classical" testing problems of validity, reliability, backwash effects, marking principles, and use of test results.

2. Why do we give tests?

Tests are often given, quite naturally, as part of the language course without too much consideration of questions like: In whose interest do we give this test? and What do we want to know something about? I would suggest that tests can be given for the following ten purposes.

- | | |
|---------------------------------|---|
| 1. Diagnosis | what do the students not know? |
| 2. Resources | who needs most help? |
| 3. Selection | who knows most? |
| 4. Prognosis/Guidance | who is most likely to succeed? |
| 5. Examination | do the students know what they should know? |
| 6. Grading | who knows most/the least etc? |
| 7. Evaluation | what have we taught the students/
have they learnt? |
| 8. Influence
(Backwash) | what the students should know and how,
what the teachers should test and how |
| 9. Stimulation
(Information) | what they know/do not know - but should
know |
| 10. Research | what is the object of study, comparisons
etc |

Tests produced for one purpose can not, as a rule, be used for a different one. This is often overlooked. When the validity of a test is commented on, one must know why it was constructed the way it was. What might be quite a good test for one purpose can be invalid for another. The term "diagnostic test" is quite often, in my experience, misapplied to almost any test. Diagnostic tests are primarily meant to find out what the students do not know, what needs to be taught or practised more.

Tests to determine the distribution of resources are used in Sweden in the teaching of Swedish as a second language to immigrant students. There are limited resources available and one needs to know who can join a regular class and who needs special teaching etc. This will be referred to in more detail below.

Some of the reasons for giving tests are more relevant in certain countries. In Sweden, for example, we give standardized tests on a nationwide basis for grading purposes, but they are not public examinations in the British or Dutch sense. Tests for evaluation come close to those used for research purposes. The American National Assessment and the British Assessment of Performance Unit projects are examples of evaluation studies: one wants to "take the temperature" of the conditions in the country.

Almost any test has some kind of backwash effect. The teacher who gives tests at regular intervals will tell her students, whether she realizes it or not, what she considers important, in what way the students are expected to use their knowledge etc. Nationwide tests given by some kind of examination board or other central body will similarly influence both students and teachers. This being so, care must be taken that these effects are positive rather than negative. This means that the backwash effects become not just chance influences that can not be avoided but purposes of their own. One reason (but not the only one, of course) why we consider the listening comprehension test so important in the Swedish standardized tests is that we believe teachers pay more attention to listening exercises because of this test than they would otherwise do.

In the following section I shall describe three different Swedish tests or test batteries, given on the basis of different "whys". In describing them I will make some further comments in detail on various background reasons for giving tests.

3. The Hows of Language Testing - three examples

Aspects to be considered under the heading "how" could refer to such things as test format, test content, marking principles, and use of test results. What weight do we attribute to validity as opposed to reliability, the two sometimes tending to counteract each other? What weight do we give various parts of the overall command of the language? How important is the actual content of the texts used? There may be a risk that language is considered a sufficient end in itself so that the texts become boring even though they are linguistically acceptable. To what extent do we use multiple-choice formats as opposed to open-ended items or student production on a larger scale? What principles do we use in marking essays etc? And so on, and so forth.

3.1 The "SVAN" Tests. More than 10 % of the students in Swedish schools are of immigrant background, many of them being bilingual and having Swedish as their second language. It has been found over the years that many - but, of course, by no means all - of these students meet with difficulties in many subjects because of language deficiencies. These are often detected too late, and there has been a general feeling that something ought to be done to "find" these students early in their school careers so that extra teaching efforts could be made. Within a research project at Gothenburg university, headed by the writer of this paper, a series of tests were developed a couple of years ago, the so-called SVAN tests (*svenska som andraspråk*, i.e. Swedish as a Second Language).

Why did we need these tests? There are 100,000 bilingual students in the comprehensive school. Instruction in Swedish as a Foreign Language (SFL) is compulsory for those students judged to be in need of such instruction. Resources being limited, the teachers have to find the students who need help the most. Any experienced teacher will be able to tell the beginner from an intermediate student, but we have found over the years that it is extremely difficult even for a well-trained teacher to tell the student with good fluency and pronunciation but a limited vocabulary from the one who has a truly native-like command of the full language. The superficial ease of the former tends to hide the limitations in the "depth" compared with the latter.

So we had to try to produce difficult tests - at each age level - to detect differences between bilingual students and monolingual students of the same age and cultural background. We also faced - but I am not sure that we solved - the problem of differentiating between general intelligence and linguistic proficiency: even among monolingual speakers there are great differences, of course, due to variation in general ability. We also wanted, through our tests, to influence the teaching procedures favorably, i.e. we aimed at positive backwash effects. We know that many, perhaps most, of the SFL teachers are not specialist language teachers. The tests were seen as part of an in-service training program.

There are oral tests for each of grades 1, 2, and 3, and both oral and written tests for grades 4, 6, 7, and 9 of the comprehensive school, students aged from 7 to 16. (It might be added that there are similar tests in Finnish, Serbo-Croatian, Croatian-Serbian, Spanish, and Turkish for the testing of the "home languages" of the bilingual children, as well as a set of models for the testing of other "home languages".)

The tests are based on pictures and series of pictures and on meaningful pieces of texts. The testing of an active and a passive vocabulary is an important part of the test battery. The words have been chosen from different spheres of life, the most important perhaps being the school environment, including textbooks, that we know from experience cause the bilingual child great problems.

The teachers were advised to evaluate pronunciation, vocabulary, grammar, and fluency, as compared to "a comparable (age, region, social background etc) monolingual student with Swedish as the first language". This was done at six levels from F-No difference to A-No proficiency.

The tests have been used on a national level for two years now and seem to work well. One of the main effects of using these tests has been that many students who were previously considered to have a proficiency at level F, i.e. no difference as compared to a monolingual student, have been given an E or a D. This in turn means that these students do get that extra SFL teaching that they still need rather than being put in the regular Swedish class. We shall have to wait for a number of years before we can tell whether there will be fewer problems at higher levels in the school system and whether the bilingual students do better in the more difficult lines of study.

The tests might be considered to be too "structuralist-psychometric" in that so much stress is put on vocabulary. The words are always tested in natural "communicative" contexts, however, and represent a great step forward from the all-too-common overemphasis on correct grammar. One of the main "days" of the tests being to detect the very good from the truly native-like, it has been necessary to use strict and reliable techniques. This is an example of the fact that the "days" explain the "hours" - but that, of course, does not mean that what has been done is not open to criticism and differences of opinion as to the results.

3.2 The Standardized Tests. In the Swedish school system nationwide standardized tests are given in certain subjects, among them foreign languages: English in grade 8 of the comprehensive school and English, French, and German in grade 2 of the *gymnasieskola*, corresponding to the eleventh year of school. These tests are given for one main reason: to help teachers give their students fair grades. The five-step grading system being norm-referenced (7-24-38-44-7 per cent given each grade), the students of the individual class have to be compared to all students in the country somehow. Since the tests are nationwide and set by the National Board of Education, they tend to become a concretization of the national curriculum, and the backwash effects are great and have to be taken into consideration in constructing the tests. This refers both to what is being tested and how it is done.

In the tests given in the spring term every year at *gymnasium* level there are the following parts: listening comprehension (multiple choice), reading comprehension (multiple choice), vocabulary-grammar tests (both MC and open-ended), and "integrative tests" (i.e. cloze). In French there is also a dictation test. In the last few years there has also been "free written production" (see below). The listening tests have sometimes consisted of actual, live recordings, but usually we have worked with scripted dialogues recorded by professional actors,

sounding somewhat like radio plays. This year's English test was a heated discussion in a miner's home on an on-going strike, originally prepared long before the present strike in Britain started. It was thus very topical and quite life-like. The reading tests have consisted both of longer pieces of texts with many questions and shorter ones with just one question for each extract. The "vocabulary-grammar" items are always monolingual and contextualized. Quite often the student is asked to produce a full phrase, and the technique comes close to a "notional-functional" approach or, if you will, a communicative test. In discussing different tests it is well to bear in mind that different "names" can well stand for quite similar tests, whereas the same "name" might hide great differences.

There is no oral test in the standardized test battery and the written essay (see below) is still included on an experimental basis. It has been found, however, in our own studies as well as in many others that the correlation between, for example, teachers' grades (which do include oral proficiency) and the standardized tests is quite high. What the teacher needs to know is the general level of her class and the approximate number of grades at each level. We can tell from our studies, for example, that teachers tend to give girls too high grades even though the girls do score somewhat higher on the standardized tests also. They also give students in slower streams lower grades than in more advanced streams even though they should be graded separately and not compared. Students in academic, "theoretical" lines of the *gymnazia* also tend to be treated more leniently than many of those in "practical" lines. Injustice is also sometimes found when students from different social backgrounds are compared. All these variations in grading which are not related to actual proficiency are counteracted by the national standardized tests.

The strong stress on communication-oriented techniques in the reading and listening tests as well as in the "vocabulary-grammar" parts tends to have backwash effects on the teaching and testing of languages in our schools. Translation is relatively little used and strict grammar tests with little or no communicative content are probably less frequent than they would be if there were no standardized tests. Since it is compulsory for teachers to use most of the tests, we try not to go against teacher opinions, and to a certain extent the tests are compromises, trying to "push" in certain directions, however.

These tests would not be very useful for diagnostic purposes, nor are they meant to be used for an evaluation of educational standards (even though certain conclusions can be drawn from a close scrutiny of the results). In evaluating the "hows" one should, again, bear in mind the "whys": primarily to grade students in a norm-referenced system and, secondly, to show how the National Board responsible for the nationwide curriculum wants it to be understood.

3.3 "Free written production". It is becoming increasingly common in many countries to put more stress on oral and written production even though the techniques used must be more or less subjectively scored. Why do we need such tests? Don't the ordinary written tests give a good enough picture of the levels of proficiency of our students? Even if foreign-language proficiency is not one unitary factor but

consists of a number of more or less distinct traits, are not the intercorrelations high enough and the test batteries we use varied enough to account for almost all of the variance?

Even if the answer to the above questions is, in all likelihood, "Yes", we still have the becksash effects, the influence of our tests on students and on teachers. It is no good stressing in the curriculum that the goal is, among other things, command of the spoken language and an ability to express oneself in writing - and then not testing it. Well, teachers are supposed to do this on their own. But we need to give them models, and the teachers have to know whether their own students do better at these things than other students. It is quite common for teachers to say that, true enough, their students scored a certain mean on the standardized test, but then their students are so good at speaking the language that the teacher has to use a higher mean in her final grades.

In order to meet this situation we have started, in the last few years, to produce tests to evaluate free (well, more or less so) written production. (Models for the testing of oral production have been produced in French and German, and similar tests have been used in English as well, but these have not been followed up quite as carefully and will not be discussed here.) There are two problems that have to be faced in this context: how do we make students produce, and how do we mark what they produce?

Experiments have shown that one and the same student can produce next to nothing if he is given a topic to write on that does not interest him at all, or he can produce a long and interesting tale on something that gets his imagination going. If we are primarily interested in his command of the foreign language, we must see to it that he is motivated to use so much of it that we have a sample to evaluate. We must also consider the "naturalness" of the task. For what purpose is a Swedish student ever likely to write something in English, French, or German? The obvious answer is: letters. We have tried other tasks, but various kinds of letters have proved suitable for our purposes. We had a native speaker write a fictitious letter-to-the-editor criticizing Swedish young people for bad and irritating behavior. The students were then asked to write their own letter to contradict or agree with "Mr. Brown". Many of these letters were rather heated and the writers may very well have forgotten that they were taking a test.

In French the first experiments with written production are being made this year. It was feared that French would be too difficult so that nothing would come of it. The kind of *précis* model used seems to have worked well, however. The students were given a text to read. It was then collected and they were asked to retell the story in a different tense. The results are quite promising. It should be borne in mind that while English is taken by 100 % of all students, less than 10 % take French, usually the linguistically very best ones.

To arrive at some kind of nationwide agreement as to the marking of the essays, we have produced booklets of examples. The papers are attributed to three "lots" of roughly equal size: average, below

average, above average. A group of teachers and teacher trainers read a couple of hundred randomly-selected papers and grade them. A dozen of them are included in the booklet and commented on. Reasons are given for the grades given. These include linguistic merits and mistakes as well as the treatment of the subject and the arrangement. Most examples illustrate the "borderlines" but there are also very good and very poor essays as well as "extremely average" ones. This technique has been well received by most teachers.

The grading is thus a combination of norm- and criterion-referencing, and communicative content is usually stressed more than formal correctness. Thus both the kind of task we set and also the treatment of the products constitute the backwash effects.

The standardized test in English in 1984 correlated .75 with teacher-given grades (given two months before the test) and .69 with the free written production test. The latter correlated .71 with teachers' grades. These figures indicate that there is great agreement between the different measures of student proficiency even though examples can also be found of students who score very high in the written essay test and low on the standardized test and vice versa.

4. Conclusion

I have tried to discuss the importance of analysing why we want to give a test before we start constructing it as well as of knowing why a certain test was constructed the way it was before one starts criticizing it. Tests can be made in many different ways but these "hows" should always be related to well-defined "whys". We should also remember that however important language testing is, language teaching is always more important. The most important of all, however, is language learning, for without it no testing is meaningful at all. Language tests, thus, are only means to certain ends and must never become ends in themselves. We must, on the other hand, always keep the ends in mind, which might be all the more important for us who are sometimes referred to, and perhaps refer to ourselves, as the language testing circles.

BEST COPY AVAILABLE